Time Series Decomposition using Automatic Learning Techniques for Predictive Models

Time Series Decomposition using Automatic Learning Techniques for Predictive Models

Jesús Silva[1], Hugo Hernández Palma[2], William Niebles Núñez[3], David Ovallos-Gazabon[4] and Noel Varela[5]

[1]Universidad Peruana de Ciencias Aplicadas, Lima, Perú.
[2] Universidad del Atlántico, Puerto Colombia, Atlántico, Colombia.
[3]Universidad de Sucre, Sincelejo, Sucre, Colombia.
[4]Universidad Simón Bolívar, Barranquilla, Atlántico, Colombia
[5]Universidad de la Costa, Barranquilla, Atlántico, Colombia.

[1]Email: jesussilvaUPC@gmail.com

Abstract. This paper proposes an innovative way to address real cases of production prediction. This approach consists in the decomposition of original time series into time sub-series according to a group of factors in order to generate a predictive model from the partial predictive models of the sub-series. The adjustment of the models is carried out by means of a set of statistic techniques and Automatic Learning. This method was compared to an intuitive method consisting of a direct prediction of time series. The results show that this approach achieves better predictive performance than the direct way, so applying a decomposition method is more appropriate for this problem than non-decomposition. The agricultural sector will be used as the study subject.

1. Introduction

Agriculture is a very important economic activity in practically all the countries of the world. In the last years, the improvement of agricultural production (quantity of production obtained by cultivated area) was caused by advances in the machinery, new techniques of sowing, and improvements in seeds and agrochemical solutions for a better control of plagues and diseases. But now, commercial agriculture has become a high technology activity in which advances in computer technology are also applied since they allow to generate high quality information on production processes [1], [2].

Therefore, agriculture can benefit from the rise of techniques within the scope of Artificial Intelligence and Data Analysis to create predictive models that predict future situations that are helpful to improve both crop productivity and decision making in all matters relating to them. Within the field of agriculture, a large number of processes such as the production or occurrence of plagues, among others, can be modeled as a collection of observations usually ordered over a period of time, that is, through a time series. By means of the application of a set of statistic techniques and Automatic Learning on time series, predictive models can be generated for extracting their regularities and making predictions [3], [4], [5].

These predictive models resulting from the data analysis can be of great importance in the cooperatives and associations on which small farmers depend when making decisions in a large number of daily situations related, for example, to plague management, crops in the field, or the management of the expected production either to offer it in the market, to hire the necessary personnel to treat it or to obtain the needed material for its preparation and packaging. Efficiency in all these situations will depend on the ability to transform raw data into accurate information that allows the right decision to be

made in each of them, with production management being one of the most important activities. Therefore, obtaining reliable information about production expectations is a critical element for the agricultural sector [6], [7], [8].

This paper decomposes the time series of the crop yield in a set of time subseries depending on a group of factors, considering the planting week, with the intention of obtaining subseries with more established patterns than the original general series. The analysis and adjustment of the predictive models of the time subseries is carried out by means of a group of statistic methods such as the ARIMA [9], and Automatic Learning methods such as neural networks among others, for making a general prediction of the crop kilograms. The proposed method was evaluated in several real case studies, and compared to a direct method consisting of a prediction of the undecomposed time series. The results show to obtain an improvement in the prediction when applying the proposed method with respect to the series without decomposing.

2. Previews Studies

In the field of agriculture, researchers have proposed several models and procedures to improve crop prediction. Most of them present a multivariate approach incorporating the impact of variables related to crops, such as rainfall, temperature, fertilizers or soil quality. The most widely used methods to perform this type of multivariate models are neural networks, more specifically in its feedforward variant. In [10] and [11], studies are developed on whether neural networks with culture-related variables are adequate to make future predictions, evaluate the predictive capacity of different parameters of the network, and compare the results with other regression models. In other studies, such as in [12], a specific study was carried out on the optimal variables for the corn harvest prediction. There is a lack of studies in terms of predicting crops with a multivariate approach and that use methods other than neural networks. Some of these studies are [13] and [14], which demonstrate the capability of Random Forests (RF) methods for crop estimation, and [15] for vector support machines. In [16], a comparison is made between different regressive methods such as neural networks, radial base functions or regression trees.

As for a univariate approach, ARIMA methods are the most widely used in crop prediction. In [17], ARIMA methods are used to predict the area and production of wheat in the coming years. Exponential smoothing studies are also applied, such as in [18]. In [19], a comparison is made on the predictive capacity between ARIMA and exponential smoothing methods.

Most of the studies carried out in this field present a multivariate point of view, focusing on the importance of the impact of endogenous factors on the variable to be predicted. The neural networks and ARIMA are the most used methods, also applying the vector support machines or the Random Forest methods, although to a lesser extent. Fewer researches have used a univariate approach that works directly on the time series.

Even so, the studies deal with time series in a very superficial way since they do not try to adapt or modify it to obtain an easier-to-predict series, but they directly focus on applying a concrete method, with different inputs or different parameterizations to obtain a better prediction. In addition, the studies just use one method for the analysis of the series, obviating the possibility of combining several methods on the same series through its subdivision.

3. Data and Methods

The problem lies within the scope of a fruit and vegetable cooperative. The cooperative is composed of a set of plots belonging to associated farmers, and production is divided into campaigns representing the crop years from September to July. Each plot plants a product during a campaign. The cultivation area of the plots is measured in squared meters and each plot has its own area. The products planted in the cooperative are different varieties of fruits and vegetables and their production is measured in kilograms. The total production of a product is formed by the sum of the production of each plot that plants the product. Some of the products are planted for a certain period of time, following a similar sowing pattern in each campaign, while others are planted irregularly throughout the campaign. The production during the life cycle of a crop is usually similar for the same product, so for products planted in nearby periods, the production follows a similar pattern.

The main interest of the cooperative is to know the production in units of weight (kilograms) that will produce each product for next week, so that the cooperative can properly manage the production volume. Therefore, the best way to measure production is on a weekly basis, and the problem to be addressed is the prediction of production in kilograms of product for the next week. This problem will be tackled by breaking down the production time series into time subseries of significant sowing weeks, and using Artificial Intelligence techniques and statistic methods to generate predictive models whose predictions will be added to finally obtain global production.

3.1 Decomposition of time series

The main idea of the time subseries decomposition approach is to capture the similar behavior that the production of a product that has been planted in the same short period of time should have, since its life cycle will be very similar. Therefore, the main decomposition criterion is taken from one of the most influential factors in production: the sowing date. Planting dates are grouped by weeks. Decomposition takes place in significant sowing weeks and there will be as many time subseries as significant planting weeks are established. Significant weeks are those that present a large number of plots in planting and, therefore, most of the production of the product. For determining the set of significant weeks, the frequencies of the sowing weeks during the previous campaigns are computed to analyze the number of plots planted in each week, and therefore choose the weeks with a greater frequency of plots in planting. Weeks with small sowing frequencies are grouped into one or two additional time series.

For the particular cases under study, decomposition was done as expressed in Table 1, which reveals the choice of significant weeks for the products, and therefore the number of time subseries.

3.2 Method

The method proposed to develop the predictive models is based on the time series decomposition on the crops yield for a product in time subseries. In each time subseries, only the crops that have been planted in a very short period of time are treated, so as an essential step for this method, the planting weeks must be analyzed and grouped to take advantage of the homogeneity in the duration and production of the crop and also cover as many plots as possible [20]. As a result, the time subseries obtained will present more predictable behaviors.

Table 1. Selected Planting Weeks for each SubSeries

No. of subseries	**Weeks of subseries planting**	
	Product 1	**Product 2**
Sub1	53	1011
Sub2	59	1112
Sub3	60	3233
Sub4		3435
Sub5		3538
Sub6		3839
Sub7		Between 14 and 33
Sub8		Not between 14 and 33

Once a partial subseries of a product is obtained, a predictive model is adjusted for each one. The predictive model building process is guided by the following steps:

1. Preprocessing: An imputation of missing values is made in the temporary subseries if they exist, so that all the time subseries present the same length. In addition, a minmax normalization is performed within the range [0.1].
2. Modeling: Predictive models of the training subseries are adjusted using a combination of statistic and machine learning techniques (ARIMA, exponential smoothing, vector support

machines, random forest, neural networks, partially recurrent neutral networks and additive models). For each of the above techniques, a set of models is generated depending on the different parameters used.

3. Validation: The validation step first determines the most suitable model for each technique and, from there, the best global model. The best model will be the one that obtains a lower error measurement and therefore a higher predictive capacity. In order to evaluate the error and predictive capacity of each of the generated models, a leave-one-out [7] cross validation is applied to the test time subseries. For this variant of cross validation, there is a single test value for each iteration, with the particularity that the training set is formed by the values that temporarily occur before the test value. Therefore, no future values to the test value are used in model training. In each iteration, the test value from the previous iteration is incorporated into the training set, and the next time value that has not been previously used is incorporated as a test value.
4. Prediction: After completing the above steps, the most suitable predictive model is obtained for each time subseries. From these models, a prediction is generated for the following week of the crop yield in each time subseries. The prediction in kilograms is obtained by multiplying the expected yield obtained from the harvest and the estimated cultivation area of the plots.
5. Aggregation: To obtain the global prediction of kilograms, the predicted values of kilograms of each time subseries are added together.

4. Results and discussion

To test validity, the approach proposed in this paper has been compared with a direct predictive approach to the problem. This direct strategy consists of applying some prediction methods to the complete time series without any decomposition. The different prediction techniques used are widely used in the current state of the art. In particular, the models considered are: ARIMA, neural networks, vector support machines, exponential smoothing, partially recurrent neural networks, Random Forest and the Facebook Prophet additive model [17]. The assumption is that this direct approach obtains less predictive ability, and therefore, a more effective approach is to consider decomposition based on the sowing week:

- Product 1. Seasonal plant. The yield production for this product is available from September 2016 to May 2018, covering the 2016, 2017, 2018 campaign. The main planting takes place in September, but small plantations are also made in the months of October and February.
- Product 2. Seasonal fruit. The yield production for this product is available from January 2016 to May 2018, covering half of the 2016, 2017 and 2018 seasons. The main planting occurs in the months of August-September and February-April.

Table 2 shows the RMSE in test obtained with the proposed method and with the direct approach. Table 3 shows the MAE. In the rows of the first column of both tables, the applied method is shown, either the one proposed in this study or the direct method. The second column specifies the modeling method used. In the case of the proposed method, a combination of all the methods have been used to adjust the series, but in the case of the direct method only one method has been used in each case. Column 3 is subdivided into two columns, each one referring to the error made in each product. All RMSE and MAE values are normalized in the range [0-1].

The results of Table 2 and Table 3 show how this approach achieves better predictive capability than the direct approach for each of the real cases by obtaining a lower error in both RMSE and MAE.

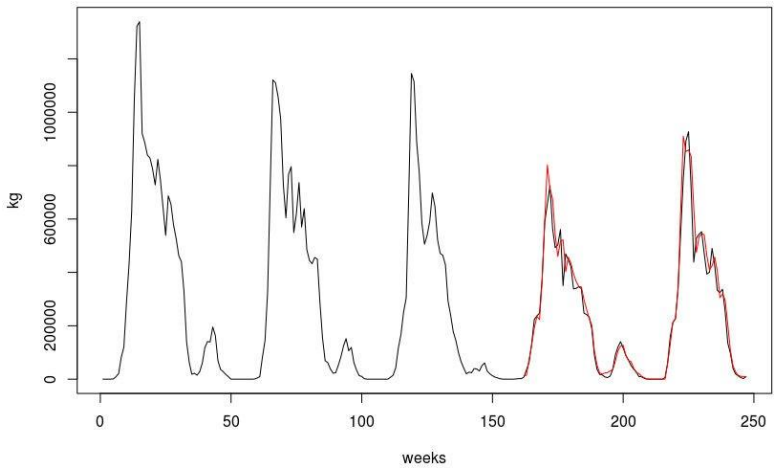
The improvement in using the decomposition approach is considerable for product 1. The RMSE obtained with this method is 0.0254, while the lowest RMSE obtained with direct method in any of its variants is 0.0511 achieved by applying a partially recurrent neural network. For the adjustments with the rest of methods in the direct method, the RMSE is superior to 0.05 and even to 0.06. On the other hand, this method obtains a lower MAE of 0.0268, while the lower MAE obtained with the direct method is 0.0299.

Table 2. Standardized RMSE obtained by the proposal and directly on the test subset.

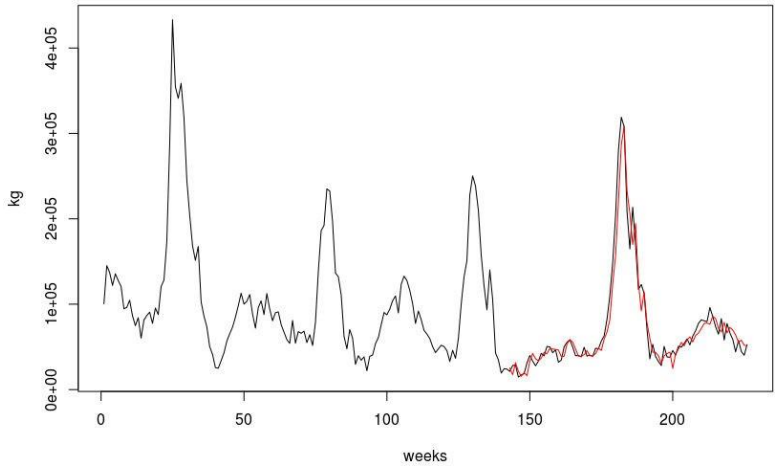
Method	**Used Model**	**Error made (RMSE)**	
		Product 1	**Product 2**
Proposed	All	0.0254	0.0289
Direct	ARIMA	0.0536	0.0472
Direct	Neural Network	0.0514	0.0524
Direct	SVM	0.0785	0.0852
Direct	Exponential Smoothing	0.0698	0.0587
Direct	Neural Network Partially Recurrent	0.0511	0.0365
Direct	Random Forest	0.0895	0.0547
Direct	Additive Model	0.0611	0.0687

While this method also offers a better predictive capability for product 2 than the direct method, the lowest RMSE obtained with this approach is 0.0289 while the lowest of the RMSE obtained by direct approach is 0.0365, also achieved with a partially recurrent neural network. The rest of the variants of the direct approach obtain a MAE greater than 0.03.

Thus, since these results show that for both product 1 and product 2 the errors made by this approach are always less than the errors obtained by the direct approach, the improvement obtained was considerable (see Figures 1 and 2).



a) Product 1



b) Product 2

Figure 1. Time series of real production and prediction in kilograms carried out by this method (red line) in test cases for the two analyzed products.

5. Conclusions

This study presents an innovative way to perform the time series analysis of the performance of production of an agricultural cooperative through the decomposition of the general time series in time subseries according to the planting week, resulting the final predictive model from the mixture between the best models adjusted for each subseries from a set of statistic methods and Automatic Learning. In addition, this method was compared with another possible implementation of the same problem.

The results of the experiments show that the proposed method improves the predictive capacity of the compared method as it obtains a minor error for both analyzed products. In addition, it always improves the error obtained in each of the comparisons. It therefore concludes that a decomposition of time series according to the significant planting weeks results in a significant improvement in the prediction over the undecomposed time series.

References

[1] Departamento Administrativo Nacional de Estadística -DANE-. (2019). Importaciones colombianas. https://www.dane.gov.co/index.php/ comercio-exterior/importaciones

[2] Jain, Mugdha, and Chakradhar Verma. "Adapting k-means for Clustering in Big Data." International Journal of Computer Applications 101.1 (2014): 19-24.

[3] Comisión Económica para América Latina y el Caribe -CEPAL-. (2013). Visión agrícola del TLC entre Colombia y Estados Unidos: preparación, negociación, implementación y aprovechamiento.

[4] Henao-Rodríguez, C., Lis-Gutiérrez, J. P., Gaitán-Angulo, M., Malagón, L. E., & Viloria, A. (2018, May). Econometric analysis of the industrial growth determinants in Colombia. In Australasian Database Conference (pp. 316-321). Springer, Cham.

[5] Lis-Gutiérrez JP., Gaitán-Angulo M., Henao L.C., Viloria A., Aguilera-Hernández D., Portillo-Medina R. (2018) Measures of Concentration and Stability: Two Pedagogical Tools for Industrial Organization Courses. In: Tan Y., Shi Y., Tang Q. (eds) Advances in Swarm Intelligence. ICSI 2018. Lecture Notes in Computer Science, vol 10942. Springer, Cham

[6] Viloria, A. "Commercial strategies providers pharmaceutical chains for logistics cost reduction." Indian Journal of Science and Technology 8, no. 1 (2016).

[7] Viloria, A., & Gaitan-Angulo, M. (2016). Statistical Adjustment Module Advanced Optimizer Planner and SAP Generated the Case of a Food Production Company. Indian Journal Of Science And Technology, 9(47). doi:10.17485/ijst/2016/v9i47/107371.
[8] J. C. Sanclemente, "Las ventas y el mercadeo, actividades indisociables y de gran impacto social y económico.: El aporte de Tosdal", Innovar, vol. 17, núm. 30, pp. 160–162, jul. 2007.
[9] N. Sapankevych y R. Sankar, "Time Series Prediction Using Support Vector Machines: A Survey", IEEE Computational Intelligence Magazine, vol. 4, núm. 2, pp. 24–38, may 2009.
[10] N. Swanson y H. White, "Forecasting economic time series using flexible versus fixed specification and linear versus nonlinear econometric models", International Journal of Forecasting, vol. 13, núm. 4, pp. 439–461, 1997.
[11] E. M. Toro, D. A. Mejia, y H. Salazar, "Pronóstico de ventas usando redes neuronales", Scientia et technica, vol. 10, núm. 26, 2004.
[12] F. Villada, N. Muñoz, y E. García, Aplicación de las Redes Neuronales al Pronóstico de Precios en Mercado de Valores, Información tecnológica, vol. 23, núm. 4, pp. 11–20. 2012.
[13] Akram, M., Bhatti, I., Ashfaq, M., Khan, A.A. Hierarchical Forecasts of Agronomy-Based Data, American Journal of Mathematical and Ma- nagement Sciences, 36(1), 49-65, 2017.
[14] Brdar S., Culibrk D., Marinkovic B., Crnobarac J., Crnojevic V. Support Vector Machines with Features Contribution Analysis for Agricultural Yield Prediction, Second International Workshop on Sensing Technolo- gies in Agriculture, Forestry and Environment, 43-47, 2011
[15] Choudhury, A. and Jones, J. Crop yield prediction using time series models, Journal of Economics and Economic Education Research., 15, 53-68, 2014.
[16] Fukuda S., Spreer W., Yasunaga E., Yuge K., Sardsud V. and Muller J. Random Forests modelling for the estimation of mango (Mangifera indica L. cv. Chok Anan) fruit yields under different irrigation regimes, Agricultural Water Management, 116(1), 142-150,2013.
[16] Ruß G. Data Mining of Agricultural Yield Data: A Comparison of Regression Models, In: Perner P. (eds) Advances in Data Mining. Applications and Theoretical Aspects, ICDM 2009. Lecture Notes in Computer Science, vol 5633.
[17] Taylor S. and Letham B. prophet: Automatic Forecasting Procedure. R package version 0.1. 2017
[18] Wuo W., Xue H. An incorporative statistic and neural approach for crop yield modelling and forecasting, Neural Computing and Applications, 21(1): 109–117, 2012.
[19] Ji, B., Sun Y., Yang S. and Wan J. Artificial neural networks for rice yield prediction in mountainous regions, Journal of Agricultural Science, 145: 249-26, 2007.
[20] Karatzoglou A., Smola A., Hornik K. and Zeileis A. kernlab - An S4 Package for Kernel Methods in R. Journal of Statistical Software, 11(9), 1-20, 2004